Amplification and adaptation of centromeric repeats in polyploid switchgrass species

Xueming Yang[1,2,*], Hainan Zhao[1,3,*], Tao Zhang[1,4], Zixian Zeng[1,3], Pingdong Zhang[1,5], Bo Zhu[1], Yonghua Han[1,6], Guilherme T. Braz[1,7], Michael D. Casler[8], Jeremy Schmutz[9,10], and Jiming Jiang[1,3,**]

[1] Department of Horticulture, University of Wisconsin-Madison, Madison, WI 53706, USA
[2] Institute of Food Crops, Provincial Key Laboratory of Agrobiology, Jiangsu Academy of Agricultural Sciences, Nanjing 210014, China
[3] Departments of Plant Biology and Horticulture, Michigan State University, East Lansing MI 48824
[4] Key Laboratory of Crop Genetics and Physiology of Jiangsu Province/Key Laboratory of Plant Functional Genomics of Ministry of Education, Yangzhou University, Yangzhou, China
[5] College of Bioscience and Biotechnology, Beijing Forestry University, Beijing, China
[6] School of Life Sciences, Jiangsu Normal University, Xuzhou, China
[7] Departmento de Biologia, Universidade Federal de Lavras, Lavras MG 37200, Brazil
[8] Dairy Forage Research Center, Agricultural Research Service, USDA, Madison, WI 53706, USA
[9] Joint Genome Institute, Walnut Creek, CA 94598, USA
[10] HudsonAlpha Institute for Biotechnology, Huntsville, AL 35806, USA

*These authors contributed equally to this work.
**Address correspondence to: jiangjm@msu.edu

Summary

- Centromeres in most higher eukaryotes are composed of long arrays of satellite repeats from a single satellite repeat family. Why centromeres are dominated by a single satellite repeat and how the satellite repeats originate and evolve are among the most intriguing and long standing questions in centromere biology.
- We identified eight satellite repeats in the centromeres of tetraploid switchgrass (*Panicum virgatum*). Seven repeats showed characteristics associated with classical centromeric repeats with monomeric lengths ranging from 166 to 187 bp. Interestingly, these repeats share an 80-bp DNA motif. We demonstrate that this 80-bp motif may dictate translational and rotational phasing of the centromeric repeats with the cenH3 nucleosomes.
- The sequence of the last centromeric repeat, Pv156, is identical to the 5S ribosomal RNA genes. We demonstrate that a 5S ribosomal RNA gene array was recruited to be the functional centromere for one of the switchgrass chromosomes.
- Our findings reveal that certain type of satellite repeats, which are associated with unique sequence features and are composed of monomers in mono-nucleosomal length, are favorable for centromeres. Centromeric repeats may undergo dynamic amplification and adaptation before the centromeres in the same species become dominated by the best adapted satellite repeat.

Key words: Centromere, cenH3 nucleosome, satellite repeats, centromere evolution, switchgrass

Introduction

Centromeres in most higher eukaryotes are composed of long arrays of satellite repeats (Henikoff *et al.*, 2001; Jiang *et al.*, 2003). In addition, centromeres in the same plant or animal species are often dominated by a single satellite repeat family. For example, human centromeres are composed exclusively of the ~171-bp alpha satellite repeats (Willard & Waye, 1987; Miga *et al.*, 2014). Similarly, each of the five centromeres of the model plant *Arabidopsis thaliana* (2n=2x=10) contains several megabases (Mb) of a 178-bp satellite repeat (Maluszynska & Heslop-Harrison, 1991; Murata *et al.*, 1994; Jackson *et al.*, 1998; Nagaki *et al.*, 2003).

Nucleosomes in centromeres are defined by the presence of cenH3 (CENP-A in mammalian

species), a centromere-specific H3 variant. The satellite repeats in a single centromere often expand to several megabases and are associated with both cenH3 nucleosomes and pericentromeric H3 nucleosomes (Schueler *et al.*, 2001; Jin *et al.*, 2004; Shibata & Murata, 2004; Zhang *et al.*, 2008). Nevertheless, such satellite repeats are intriguingly restricted to the centromeric regions and do not spread to interstitial or telomeric regions of the chromosomes.

Centromeric satellite repeats can evolve rapidly. For example, centromeres of rice (*Oryza sativa*) chromosomes contain a 155-bp satellite repeat CentO (Cheng *et al.*, 2002). The CentO repeats, however, were lost or replaced by different centromeric satellite repeats in several closely related *Oryza* species (Lee *et al.*, 2005; Yi *et al.*, 2013; Yi *et al.*, 2015). It has long been recognized that the monomeric lengths of many centromeric satellite repeats range from 150 to 180 bp, which is desirable for wrapping a single nucleosome (Jiang *et al.*, 2003). This “one monomer for one cenH3 nucleosome” perception has recently been confirmed experimentally in both humans (Hasson *et al.*, 2013) and rice (Zhang *et al.*, 2013). Centromeric satellite repeats are highly phased with the cenH3 nucleosomes (following the “one repeat monomer for one cenH3 nucleosome” pattern or periodicity)

in both human and rice, and is considered to be favorable to the stability of the cenH3 nucleosomes (Zhang *et al.*, 2013).

Despite of our understanding on why satellite repeats are favorable DNA sequences for cenH3 nucleosomes, the origin and evolution of the centromeric satellite repeats remains elusive. Interestingly, not all centromeres contain satellite repeats. Centromeres can be activated from non-centromeric regions that lack repetitive DNA sequences (Nasuda *et al.*, 2005; Topp *et al.*, 2009; Fu *et al.*, 2013; Wang *et al.*, 2014; Liu *et al.*, 2015). Native centromeres that lack satellite repeats were also found in several plant and animal species (Shang *et al.*, 2010; Gong *et al.*,

2012; Purgato *et al.*, 2015). It was hypothesized that such repeat-free centromeres will eventually be invaded by satellite repeats and transformed as the classical satellite repeat based centromeres (Yan *et al.*, 2006; Gong *et al.*, 2012). Nevertheless, there is only limited information about the evolutionary process pushing repeat-free centromeres towards centromeres dominated by a single satellite repeat.

We conducted a genome-wide characterization of DNA sequences associated with cenH3 nucleosomes in switchgrass (*Panicum virgatum*) (2n=4x=36), an allotetraploid species with KKNN genomes. We identified eight satellite repeats associated with different centromeres. One of the repeats was found to be identical with the 5S ribosomal RNA genes (5S rDNA) in switchgrass. We demonstrate that a 5S rDNA array was adapted to the functional centromere of one switchgrass chromosome. Surprisingly, the remaining seven repeats shared an 80-bp DNA motif, suggesting that these repeats were originated from the same ancestral repeat or repeat family. We demonstrate that this 80-bp motif likely dictates translational and rotational phasing of the centromeric repeats with the cenH3 nucleosomes, and may be favorable to the assembly and maintenance of cenH3 nucleosomes and confer stability to centromeric chromatin.

Materials and Methods

FISH and ChIP-seq

Tetraploid switchgrass cultivars Kanlow and Summer were used in immunofluorescence and fluorescence in situ hybridization (FISH) analyses. Both Kanlow and Summer were used in mapping centromeric repeats and these cultivars showed similar chromosomal distribution patterns of all repeats. An octoploid cultivar Trailblazer was used in FISH mapping. Root tips were harvested from plants grown in the green house and fixed using 3:1 ethanol:glacial acetic acid. FISH was performed following published procedures using regular hybridization and washing conditions (Jiang *et al.*, 1996). DNA probes were labeled by digoxigenin-11-dUTP or biotin-16-dUTP.

The hybridization signals were detected with Alexafluor 488 streptavidin for biotin-labeled probes, and rhodamine-conjugated anti-digoxigenin for dig-labeled probe.
Chromosomes were counterstained with 4′,6-diamidino-2-phenylindole (DAPI) in VECTASHIELD antifade mount media (Vector Laboratories, Burlingame, CA). FISH images were captured using a QImaging Retiga EXi Fast 1394 CCD camera attached to an Olympus BX51 epifluorescence microscope. Images were processed with Meta Imaging Series 7.5

software. The final contrast of the images was processed using Adobe Photoshop CS3 software. A rice cenH3 antibody (Nagaki *et al.*, 2004) was used for chromatin immunoprecipitation (ChIP). ChIP, ChIP followed by Illumina sequencing (ChIP-seq), and mapping of ChIP-seq reads were performed following published protocols (Nagaki *et al.*, 2003; Zhang *et al.*, 2012).

Annotation of switchgrass satellite repeats

Repetitive DNA sequences in the switchgrass genome were identified by RepeatExplorer (Novak *et al.*, 2013) using 10 millions of random shotgun reads (250 bp). To obtain the consensus sequences of the centromeric repeats, we mapped cenH3 ChIP-seq reads iteratively to the output sequences of RepeatExplorer. Fixed consensus sequences of the repeats were obtained after the three rounds of iterative mapping. We then used the consensus sequences to search PacBio reads associated with each repeat by BLAST with parameters '-task blastn' (Zhang *et al.*, 2000). A new set of consensus sequences were constructed using the monomers identified in the PacBio reads. The two consensus sequences constructed using these two different methods were identical for each repeat. The cenH3 enrichment for each repeat was determined as described in previous reports (Gong *et al.*, 2012; Zhang *et al.*, 2014). The annotation of Pv1, Pv2, Pv29, Pv36, Pv45, Pv115, Pv118 and Pv156 were conducted using Repbase (Kohany *et al.*, 2006). The coding and spacer regions were determined by searching against 5S rRNA gene sequences in grass species (Szymanski *et al.*, 2002). The assembly of 5S rDNA associated reads were conducted by CAP3 (Huang & Madan, 1999) with parameters "-o 40 –p 80". To estimate the amount of 5S rDNA in the genome, we mapped genomic sequence reads (SRR387527 and SRR387530) generated by DOE joint genome institute (JGI) to the full length of 5S rRNA gene. The estimated amount of 5S rDNA was calculated by multiple the genome size (1.6 Gb) to the percentage of reads which can be mapped to the 5S rRNA gene. To obtain the full length of the repeat unit (the monomer) of each repeat, cenH3 ChIP-seq reads were aligned to each repeat

using BWA with default parameters (Li & Durbin, 2009). The 5’ end and 3’ end were then extended until the monomer was fully covered. Multiple sequence alignment of monomers were conducted by Clustal Omega with default setting (Sievers *et al.*, 2011).

Nucleosome occupancy and WW dimer periodicity associated with each repeat.

ChIP-seq read pairs were merged using FLASH2 with parameters “-m 5 –M 1000 –e 35 – x 0.3” (Magoc & Salzberg, 2011). The merged fragments were aligned to a tetramer (four copies of the monomer) of each repeat using BWA with default parameters (Li & Durbin, 2009). To analyze the association of cenH3 nucleosomes with each repeat, we mapped ChIP-seq DNA fragments to the seven Pv repeats and assigned each fragment to one of the seven repeats based on its best mapping quality. For fragments with the same mapping quality to multiple repeats, we randomly assigned these fragments to one of these repeats. Fragments were divided in to two groups, small fragments (less than 130 bp) and large fragments (130-170 bp). The frequencies of midpoints along the tetramers were calculated. The nucleosome occupancy were analyzed by plotting the fragment lengths against the positions on a tetramer of each repeat. We analyzed the frequency and location of dinucleotide SS (G/C) and WW (A/T) associated with the centromeric repeats. The WW dimer periodicity, which is defined as the distribution of the distances between two WW dimers, was calculated as described in a previous report (Zhang *et al.*, 2013). To calculate the phasing score, the distances between WW dimers on target sequence were calculated. The odds of 9.8 bp WW dimer O_i were calculated:

$$O_i = \frac{f_i}{\frac{n_i}{L}}$$

L is the length of target sequence, f_i is the frequency of WW dimer. n_i is the number of WW dimer with distance i. i belong to {9, 10, 19, 20… $n10$-1, $n10$} and $n10 < L$. The phasing score is the log2 of the median of the odds on the target sequence.

Results

DNA sequences associated with cenH3 nucleosomes in switchgrass

We performed immunofluorescence assays using an anti-cenH3 antibody developed in rice (Nagaki *et al.*, 2004). This antibody was found to specifically label the centromeres in distantly related grass species,

including wheat, maize, and oat (Jin *et al.*, 2004; Nasuda *et al.*, 2005; Zhang *et al.*, 2010). Centromere-specific immunofluorescence signals were detected on all switchgrass chromosomes (**Figures 1A**). We then conducted ChIP using nuclei isolated from leaf tissue of switchgrass cultivar Summer. DNA from ChIP was labeled as a probe for FISH. We detected FISH signals in the centromeres of most switchgrass chromosomes (**Figure 1B**).

However, the sizes and intensities of the FISH signals varied significantly among the

centromeres. If a centromere contains highly repetitive satellite repeats, it would show strong FISH signals from a ChIPed DNA probe. By contrast, if a centromere contains mainly single or low copy sequences, it would show weak or no FISH signals from a ChIPed DNA probe (Gong *et al.*, 2012; Zhang *et al.*, 2014). Thus, the FISH results from the ChIPed DNA probe suggested that switchgrass centromeres contain variable amounts of repetitive DNA sequences and some centromeres may contain single or/and low copy sequences.

A ChIP DNA library and an input DNA library were developed and sequenced using the Illumina HiSeq platform. We generated 32.1 and 33.7 millions of 100-bp paired-end sequence reads from the two libraries. Approximately 71% of the paired-end reads from the ChIP-seq library were mapped to a unique position in the assembled *P. virgatum* v4.1 genome (https://phytozome.jgi.doe.gov/pz/portal.html#!info?alias=Org_Pvirgatum_er). The relative enrichment of the ChIP-seq reads in the genome was normalized using sequence reads from the input library. The distribution of unique ChIP-seq reads was displayed in 50-kb windows along the 18 switchgrass chromosomes (**Figure 2**, **Figure S1**). Read-enriched regions were observed in several chromosomes. For example, chromosome 2K included a read-enriched domain spanning ~2 Mb. By contrast, no sequence enrichment was observed in chromosome 8K (**Figure 2**). Several switchgrass chromosomes contained multiple read-enriched regions throughout the chromosomes (**Figure S1**), which is likely caused by mis-assembly of the centromeric sequences and/or mapping artifacts caused by repetitive DNA sequences. Because most switchgrass chromosomes lack a single megabase-sized cenH3 read-enriched domain, we conclude that most switchgrass centromeres contain mainly or exclusively repetitive DNA sequences, which is a common feature for the centromeres in most plant species (Jiang *et al.*, 2003).

Computational identification of repeats enriched in the centromeres

We intended to identify the centromere-specific repeat(s) in the switchgrass genome. We first used 10 millions of random shotgun reads (250 bp) to computationally identify all repetitive DNA sequence clusters using a similarity-based sequence clustering approach (Macas *et al.*, 2007; Novak *et al.*, 2010). We identified a total of 64,194 repeat clusters (CL). The sequence proportion (%) of each cluster can be estimated based on the number of reads associated with the cluster. We then mapped the cenH3 ChIP-seq reads to the clusters and calculated ratios of ChIP- seq reads to shotgun reads associated with each cluster (**Table 1**). This ChIP/shotgun ratio is

indicative to the relative enrichment of each repeat in the centromeres (Gong *et al.*, 2012; Neumann *et al.*, 2012; Kowar *et al.*, 2016). We plotted the ChIP/ shotgun ratios of 200 most abundant repeat clusters. Most clusters showed a ratio close to 1, indicating that these repeats are not enriched in the centromeres. By contrast, eight clusters showed a ratio >3 (**Figure 3A, Table 1**), suggesting that these repeat clusters are likely associated with the centromeres. Most strikingly, CL156 showed nearly 25 fold of enrichment in the centromeres. CL1 accounted 1.6% of the shotgun reads, representing the most abundant repeat family in the switchgrass genome.

Cytological confirmation of the computationally identified centromeric repeats

We next investigated if the eight repeat clusters enriched with ChIP-seq reads are truly associated with the centromeres. We designed specific primers for each of the eight clusters (**Table S1**). Polymerase chain reaction (PCR) using these primers revealed a ladder amplification pattern for several clusters (**Figure 3B**), suggesting that these clusters are associated with typical satellite repeats. One amplified DNA fragment from each cluster was cloned and several plasmids were sequenced. The genomic sequences homologous to the plasmids were extracted and used to develop a consensus sequence for each repeat. For example, a 270-bp fragment associated with CL1 (**Figure 3B**) was cloned and named the corresponding *Panicum virgatum* (Pv) repeat Pv1. Sequence analysis showed that the monomer of the Pv1 repeat is 166 bp. The monomers of all eight Pv repeat were in typical mono-nucleosomal size, ranging from 166 bp to 187 bp, except for Pv156 that was 326 bp (**Table 1, Table S2**).

To further analyze the tandem organization of these Pv repeats, we searched and identified PacBio sequence reads generated by the Switchgrass Genome Sequencing Consortium (see Materials and Methods). We identified a total of 165,589 PacBio reads containing at least one of these eight repeats. The average length of these reads is 15.4 kb, including 8,825 reads longer than 30 kb. Sequence analysis showed that many long PacBio

reads contain nearly exclusively tandem arrays of one specific Pv repeat (**Table S3**). For example, read 2802313_2295_44761_41341 is 41-kb long and contains 247 tandemly organized monomers of Pv1, which comprised of 99% of this read. Similar results were found for all eight repeats (**Table S3**). These results confirmed that these computationally identified satellite repeats are organized into tandem arrays in the switchgrass genome.

A representative plasmid clone from each Pv repeat was then used for FISH using normal

hybridization and washing condition. All eight plasmids produced FISH signals that were highly specific to centromeres. Repeat Pv36 hybridized to the centromeres of a single pair of chromosomes (**Figure 4A**). Similarly, Pv156 produced major FISH signals in the centromeres of a single pair of chromosomes, and minor signals in another pair of centromeres (**Figure 4B**).

Dual-color FISH revealed that Pv36 and Pv156 are located on different chromosomes and Pv156 is clearly more abundant than Pv36 based on the size and intensity of the FISH signals (**Figure 4C**), which matched to the results based on sequence read counts (**Figure 3A**).

Pv1 and Pv115 produced a similar FISH signal pattern. Both repeats hybridized to the centromeres of every chromosomes, but approximately half of the chromosomes showed stronger signals than the remaining chromosomes (**Figures 4D, 4E**). The remaining four repeats, Pv2, Pv29, Pv45, and Pv118, hybridized to the centromeres of all chromosomes. However, a few centromeres showed stronger signals than the rest of the centromeres. Both Pv29 (**Figure 4F**) and Pv118 (**Figure 4G**) produced strong signals in a single pair of chromosomes. However, the strong Pv29 and Pv118 signals were associated with different centromeres (**Figure S2**). Pv2 (**Figure 4H**) and Pv45 (**Figure 4I**) produced strong signals in two pairs of chromosomes. Dual- color FISH revealed that one pair of centromeres showed strong hybridization signals from both Pv2 and Pv45 (**Figure S2**). Since seven Pv repeats share sequence similarities in a 80-bp region (see below), we cannot exclude the possibility that weak signals from some Pv repeats are derived from cross-hybridization from another repeat(s).

FISH mapping of centromeric repeats in octoploid switchgrass

Phylogenetic studies showed that the octoploid switchgrass (2n=8x=72) was likely derived from hybridization between disparate tetraploids (Triplett *et al.*, 2012). However, a DNA marker-based analysis suggested that upland tetraploid arose from upland octoploid (Lu *et al.*, 2013). Nevertheless, 8x switchgrass should be genomically doubled

compared to 4x switchgrass. We conducted FISH on chromosomes from an octoploid switchgrass cultivar “Trailblazer” using several centromeric repeats. The FISH signals patterns on Trailblazer chromosomes were essentially double of the patterns observed on tetraploid switchgrass chromosomes. For examples, Pv36 was observed on four chromosomes (**Figure 5A**). Four major and four minor signals of Pv156 signals were observed on eight Trailblazer chromosomes (**Figure 5B**). Four 4x switchgrass (**Figure 4H**) and approximately eight 8x switchgrass chromosomes (**Figure 5D**)

showed strong FISH signals derived Pv2. Pv1 hybridized strongly to half of the chromosomes in both 4x and 8x switchgrass chromosomes (**Figure 4D, Figure 5C**).

5S ribosomal RNA genes were recruited as centromeric DNA

We next investigated the origin of the centromeric repeats based on sequence similarities with known repeats in the literature. Surprisingly, Pv156, the most centromere-enriched repeat (**Figure 3A**), was found to be related with the 5S ribosomal RNA genes (5S rDNA). Pv156 is 326 bp long and has an identical length as a single unit of the 5S rDNA (**Figure 6A**). Pv156 contains a sequence (49 to 167 bp) that is an equivalent of the complete coding sequence of the 5S rRNA gene. This putative coding sequence of Pv156 is flanked by spacer-related sequences (**Figure 6A**). The predicted coding region within Pv156 showed 99-100% sequence identity with the coding sequence of 5S rRNA genes from other grass species, including rice (*Oryza sativa*), maize (*Zea mays*), and wheat (*Triticum aestivum*). A satellite repeat, PLsatB, identified in *Plantago lagopus*, was found to be derived the 5S rDNA (Kumke *et al.*, 2016). However, the monomers (459-505 bp) of PLsatB contain only 82 bp sequences derived from 5S rDNA.

We found only five 5S rDNA copies (more than 80% identity and 50% coverage) associated with chromosome 8N, 5K, 1K and a non-anchored scaffold (scaffold_1075, 26,648 bp) in the *P. virgatum* v4.1 genome, suggesting that most of the 5S rDNA array(s) is not included in the assembled genome. We tried to assemble the switchgrass 5S rDNA using genomic sequences from our input DNA library and other publically available genomic sequences. We collected all genomic sequences with >70% identity with the Pv156 repeat. We obtained 4,153 DNA fragments from our input library (average 148 bp) and 11,789 of 454 sequences (average 247 bp). Our assembling process yielded many 5S rDNA arrays with different lengths. The two longest arrays spanned 14.2 kb and 11.1

kb, respectively. These two arrays contained a total of 82 5S rDNA units, including 46 intact and 36 truncated units. The coding and spacer sequences within the Pv156 repeat showed averages of 98% and 95% sequence identity to the consensus coding and spacer sequences derived from the 46 intact 5S rDNA units. We also identified 4,729 long PacBio reads containing long arrays of 5S rDNA (**Table S3**). For example, read 1451234_31_40373_39567 is 39.6 kb long and contain 112 5S rDNA monomers. The monomers from the PacBio reads had 97.1% sequence similarity with Pv156. Finally, we conducted dual-color FISH using Pv156 and a 5S rDNA probe cloned from

rice. The FISH signals derived from the two probes overlapped completely, including a pair of major signals and a pair of minor signals, both in the centromeres (**Figure 6B**). Collectively, these results showed that the Pv156 repeat is structurally not different from the 5S rDNA assembled using genomic sequences and those from PacBio sequences. Thus, the switchgrass chromosome with the major 5S rDNA array likely recruited portion of the 5S rDNA array as the functional centromere.

Centromeric satellite repeats share an 80-bp evolutionary conserved DNA motif

The remaining seven Pv repeats showed different FISH signal patterns, ranging from locations to a single pair of centromeres (Pv36) to locations to every centromere (Pv1).

Surprisingly, these seven repeats showed sequence similarities among themselves as well as to the centromeric satellite repeats CentO (155 bp) from rice (Dong *et al.*, 1998; Cheng *et al.*, 2002) and CentC (156 bp) from maize (Ananiev *et al.*, 1998) (**Figure 7, Table S4**). Previously, a conserved 80-bp motif was found in the centromeric satellite repeats from several distantly related grass species, including both CentO and CentC (Lee *et al.*, 2005). Strikingly, this 80-bp motif was found in all seven Pv repeats (**Figure 7**). These results suggest that the seven Pv repeats were likely derived from the same ancestral repeat or repeat family. Presence of this 80- bp motif in multiple centromeric repeats in switchgrass further revealed an intriguing sequence property associated with this motif, which may be especially adaptable for cenH3 nucleosomes.

The 80-bp motif represents the core sequence wrapping cenH3 nucleosomes

We next investigated how each Pv repeat wraps the cenH3 nucleosomes. In the cenH3 ChIP experiments the switchgrass chromatin was digested by *Micrococcal Nuclease* (MNase) into a size composed of mostly single nucleosomes. DNA fragments in mono-nucleosomal size from

both ChIP and input were collected and sequenced. The ChIP and input libraries were sequenced using 100 bp paired-end mode to ensure the recovery of the entire nucleosome- protected DNA sequences. Overlapping sequences from each pair of reads were merged into a single DNA fragment, which represents the DNA sequence protected by a single nucleosome. Approximately 93% of the read pairs can be merged. The lengths of merged fragments from the input library, which represent the bulk nucleosomes, showed one major peak at 148 bp. By contrast, the lengths of merged fragments from the ChIP library showed two major peaks at 147

bp and 125 bp, respectively (**Figure S3**), which is consistent with the previous reports that cenH3 nucleosomes protect less DNA sequences than canonical nucleosomes in both plant and animal species (Hasson *et al.*, 2013; Zhang *et al.*, 2013; Zhao *et al.*, 2016).

The merged fragments were mapped to a tetramer (four copies of a consensus monomer) of each Pv repeat. We divided the fragments derived from the ChIP library into small (<130 bp) and large (>130 bp) groups. The small fragments represent sequences associated with cenH3 nucleosomes; while the large fragments likely represent mix of sequences derived from bulk nucleosomes and under-digested cenH3 nucleosomes. The small fragments from several Pv repeats, including Pv1 (**Figure 8A**), Pv115 (**Figure 8B**), Pv2 and Pv29 (**Figure S4**) were translationally phased, in which the putative nucleosomes occupy specific DNA regions and the distances between the nucleosomes show periodicity (Zhang *et al.*, 2013). Interestingly, the phased fragments spanned nearly the entire 80-bp motif. Thus, there are preferred sites for MNase digestion within these repeats and the 80-bp motifs represent sequences that wrap the cores of the cenH3 nucleosomes. The large fragments from the ChIP library and fragments from the input library of Pv1 (**Figure 8A**) and Pv2 (**Figure S4**) were phased similarly as the small fragments from the ChIP library. By contrast, the large fragments from Pv115 (**Figure 8B**) and Pv29 (**Figure S4**) showed different phasing patterns and the phased fragments did not span the 80-bp motif. Therefore, the Pv115 and Pv29 repeats appeared to wrap differently for bulk nucleosomes compared to cenH3 nucleosomes.

The nucleosomal wrapping patterns of Pv45 and Pv118 were more complex. Small fragments from the ChIP libraries of Pv45 and Pv118 were not clearly phased or appeared to be associated with multiple phasing patterns (**Figure S4**). The 80-bp motif was either included or excluded in the phased fragments. The small fragments from the ChIP library of Pv36 were phased, however, the 80-bp motif was located in the middle of phased cenH3 nucleosomes (**Figure S4**).

Rotational phasing of centromeric satellite repeats with cenH3 nucleosomes

It is well known that DNA sequences, especially the frequency and location of dinucleotide SS (G/C) and WW (A/T), are important for nucleosome positioning (Segal *et al.*, 2006; Valouev *et al.*, 2008; Ioshikhes *et al.*, 2011). A periodicity of WW dinucleotide about every ~10 bp, which corresponds to one turn of the DNA double helix, is believed to favor a

particular orientation of the DNA sequence toward the nucleosome core and is known as rotational phasing of nucleosomes (Segal *et al.*, 2006; Ioshikhes *et al.*, 2011). We analyzed the periodicity of WW dimers associated with each of the seven Pv repeats. Interestingly, we observed a strong ~10-bp WW periodicity associated with the 80-bp motifs in several Pv repeats, including Pv1, Pv2, Pv29, and Pv115. In contrast, such a periodicity was much fuzzier or not visible in sequences that flank the 80-bp motif of these Pv repeats (**Figures 8C, 8D, Figure S5**). These results further support that the 80-bp motif may provide a favorable sequence feature for wrapping the cenH3 nucleosomes.

We measured the degree of ~10-bp WW dimer periodicity of each repeat by calculating the phase score (see Materials and Methods). The phase scores in the 80-bp motifs were especially high for Pv1, Pv2, Pv29, and Pv115 and were clearly higher than in regions flanking the 80-bp motif. In contrast, Pv36, Pv45 and Pv118 showed low phase scores. In addition, the 80-bp motifs of Pv45 and Pv118 showed lower phase scores than the flanking regions (**Figure S5**). Interestingly, the lack of ~10-bp WW dimer periodicity within Pv45 and Pv118 was correlated with the lack of phasing of the ChIPed reads (<130 bp) derived from these two Pv repeats (**Figure S5**).

Discussion

Centromeres in most animal and plant species are dominated by a single satellite repeat. It has been a challenge to investigate the origin of such “fully established” centromeric repeats because these repeats may have evolved over the course of millions of years. Newly emerged or “young” satellite repeats in centromeres were reported in *Solanum* species because these repeats are absent in closely related plant species (Gong *et al.*, 2012; Zhang *et al.*, 2014). Most of such young centromeric repeats were amplified from retrotransposon-related sequences (Gong *et al.*, 2012; Zhang *et al.*, 2014). Several chicken centromeres contain satellite repeats that are specific to individual centromeres. These repeats also appeared to be

amplified from retrotransposon- related sequences (Shang *et al.*, 2010). Thus, retrotransposon-related sequences are a common seeding source for the origin of satellite repeats in centromeres.

The monomeric lengths of fully established centromeric satellite repeats are often in the range of a single nucleosome, typically from 150 bp to 180 bp. Interestingly, the monomeric lengths of seven Pv repeats are all within this range, from 166 bp of Pv1 to 187 bp of Pv118

(**Table 1**). By contrast, the monomeric lengths of newly emerged centromeric repeats vary widely, ranging from a few hundred base pairs to several kilobases (Gong *et al.*, 2012; Zhang *et al.*, 2014). Centromeric satellite repeats with 'odd' monomeric lengths, which are significantly deviated from the mono-nucleosomal length, were also reported in several plant species (Lee *et al.*, 2005; Nagaki *et al.*, 2012; Neumann *et al.*, 2012; Iwata *et al.*, 2013; Melters *et al.*, 2013; Kowar *et al.*, 2016). Interestingly, these repeats were often restricted to a few centromeres and have not been spread to all centromeres (Lee *et al.*, 2005; Nagaki *et al.*, 2012; Neumann *et al.*, 2012; Iwata *et al.*, 2013). Thus, satellite repeats with monomeric lengths in single nucleosome size appear to be intriguingly suitable for centromeres. One potential advantage for such satellite repeats is that each monomer can be adapted for a single cenH3 nucleosome (Hasson *et al.*, 2013; Zhang *et al.*, 2013). Translationally phased cenH3 nucleosome arrays can be readily assembled based on such tandem repeats, which may be favorable for establishment and maintenance of the cenH3 nucleosomes.

Specific sequence features associated with a satellite repeat are likely another factor to be favorable for centromeres. Strikingly, seven Pv repeats share an 80-bp motif that is also associated with the well-studied repeats CentO and CentC (**Figure 7**). All known satellite repeats containing this motif are restricted to centromeres, suggesting that this 80-bp motif has an intriguing sequence property favorable for cenH3 nucleosome assembly (Lee *et al.*, 2005). Here we demonstrate a strong ~10-bp WW periodicity associated with this 80-bp motif in several Pv repeats (**Figure 8**). Pv1 and Pv2 are the most abundant centromeric repeats in switchgrass (**Figure 3**). The cenH3 ChIPed sequences from Pv1 and Pv2 were highly phased and the 80-bp regions were spanned by the phased sequences. These results showed that the monomers of Pv1 and Pv2 rotationally and translationally phased with the cenH3 nucleosomes with the 80-bp region wrapping the core of the nucleosomes. By contrast, the ~10-bp WW periodicity within the 80-bp regions was much weaker in Pv45 and Pv118 and the cenH3 ChIPed sequences from these two repeats were not phased. These results support

that satellite repeats with specific sequence features, such as a strong ~10-bp WW periodicity, will be favorable for cenH3 nucleosome assembly. These results suggest that multiple centromeric repeats may emerge and undergo dynamic amplification and adaptation, especially after an allopolyploidization event that combined two progenitor genomes that may contain different centromeric repeats. However, the

best-adapted satellite repeat will survive from this dynamics and will eventually occupy all centromeres in the same species.

We discovered a surprising adaptation of the 5S rDNA (Pv156) in switchgrass centromeres. A single unit of 5S rDNA in switchgrass is 326 bp (119 bp coding plus 207 bp spacer). Analysis of random genomic sequences showed that the 5S rDNA-related sequences account for 0.2% of the switchgrass genome, representing a total of 3.2 Mb sequences. Most of this 3.2 Mb array is located in a single centromere (**Figure 4B**). Similarly, estimation based on RepeatExplorer indicated that the 5S rDNA accounts for 0.24% (3.8 Mb) of the switchgrass genome. The cenH3-binding domains of switchgrass centromeres are approximately 2 Mb based on the size of ChIPed sequence-enriched domains in some switchgrass chromosomes (**Figure 2**). It is likely that only part of the 3.2-Mb 5S rDNA array was recruited as centromeric DNA. Thus, although cenH3 nucleosomes are antagonistic to transcription (Zhao *et al.*, 2016), incorporating part of this massive array into a centromere appeared to not impede the function of the 5S ribosomal RNA genes in switchgrass. The ChIPed DNA reads associated with Pv156 were not phased (**Figure S4**) and no ~10-bp WW periodicity was detected within the Pv156 repeat. Thus, Pv156 resembles a young centromeric satellite repeat with an odd monomeric length and does not assemble phased cenH3 nucleosomes.

In conclusion, our findings in switchgrass reveal new insights of the origin and evolution centromeric DNA sequences. We demonstrate that the repetitive DNA sequences in the centromeres may undergo dynamic amplification and adaptation before centromeres become dominated by the best adapted satellite repeat, which may be associated with unique sequence features and is composed of monomers in mono-nucleosomal length, are favorable for centromeres.

Availability of data and materials

The cenH3 ChIP-seq and input sequencing data is available from NCBI Sequence Read Archive (SRA) under BioProject PRJNA397205.

Acknowledgements

This work was supported partially by National Science Foundation (NSF) grant IOS-1444514 to

J.J. The work conducted by the U.S. Department of Energy Joint Genome Institute is supported

by the Office of Science of the U.S. Department of Energy under Contract No. DE-AC02- 05CH11231. We thank the JGI for prepublication access to the *P. virgatum* genome sequence.

Author contributions

J.J. conceived the research. X.M.Y., T.Z., Z.X.Z., P.D.Z., B.Z., Y.H.H., and G.T.B. performed experiments. H.N.Z., T.Z. and J.J. analyzed the data. M.D.C. and J.S provided resource. H.N.Z. and J.J. wrote the manuscript.

ORCID

Jiming Jiang: https://orcid.org/0000-0002-6435-6140 Jeremy Schmutz: http://orcid.org/0000-0001-8062-9172

Reference

Ananiev EV, Phillips RL, Rines HW. 1998. Chromosome-specific molecular organization of maize (*Zea mays* L.) centromeric regions. *Proc. Natl. Acad. Sci. USA* **95**: 13073-13078.

Cheng ZK, Dong FG, Langdon T, Ouyang S, Buell CB, Gu MH, Blattner FR, Jiang JM. 2002. Functional rice centromeres are marked by a satellite repeat and a centromere- specific retrotransposon. *Plant Cell* **14**: 1691-1704.

Dong FG, Miller JT, Jackson SA, Wang GL, Ronald PC, Jiang JM. 1998. Rice (*Oryza sativa*) centromeric regions consist of complex DNA. *Proc. Natl. Acad. Sci. USA* **95**: 8135-8140.

Fu SL, Lv ZL, Gao Z, Wu HJ, Pang JL, Zhang B, Dong QH, Guo X, Wang XJ, Birchler JA, et al. 2013. De novo centromere formation on a chromosome fragment in maize. *Proc. Natl. Acad. Sci. USA* **110**: 6033-6036.

Gong ZY, Wu YF, Koblizkova A, Torres GA, Wang K, Iovene M, Neumann P, Zhang WL, Novak P, Buell CR, et al. 2012. Repeatless and repeat-based centromeres in potato: implications for centromere evolution. *Plant Cell* **24**: 3559-3574.

Hasson D, Panchenko T, Salimian KJ, Salman MU, Sekulic N, Alonso A, Warburton PE, Black BE. 2013. The octamer is the major form of CENP-A nucleosomes at human centromeres. *Nature Structural & Molecular Biology* **20**: 687-695.

Henikoff S, Ahmad K, Malik HS. 2001. The centromere paradox: Stable inheritance with rapidly evolving DNA. *Science* **293**: 1098-1102.

Huang XQ, Madan A. 1999. CAP3: A DNA sequence assembly program.

Genome Research **9**: 868-877.

Ioshikhes I, Hosid S, Pugh BF. 2011. Variety of genomic DNA patterns for nucleosome positioning. *Genome Res.* **21**: 1863-1871.

Iwata A, Tek AL, Richard MMS, Abernathy B, Fonseca A, Schmutz J, Chen NWG, Thareau V, Magdelenat G, Li YP, et al. 2013. Identification and characterization of functional centromeres of the common bean. *Plant Journal* **76**: 47-60.

Jackson SA, Wang ML, Goodman HM, Jiang JM. 1998. Application of fiber-FISH in physical mapping of *Arabidopsis thaliana*. *Genome* **41**: 566-572.

Jiang JM, Birchler JA, Parrott WA, Dawe RK. 2003. A molecular view of plant centromeres.
Trends Plant Sci. **8**(12): 570-575.

Jiang JM, Hulbert SH, Gill BS, Ward DC. 1996. Interphase fluorescence in situ hybridization mapping: A physical mapping strategy for plant species with large complex genomes.
Molecular and General Genetics **252**: 497-502.

Jin WW, Melo JR, Nagaki K, Talbert PB, Henikoff S, Dawe RK, Jiang JM. 2004. Maize centromeres: Organization and functional adaptation in the genetic background of oat. *Plant Cell* **16**: 571-581.

Kohany O, Gentles AJ, Hankus L, Jurka J. 2006. Annotation, submission and screening of repetitive elements in Repbase: RepbaseSubmitter and Censor. *BMC Bioinformatics* **7**. **Kowar T, Zakrzewski F, Macas J, Koblizkova A, Viehoever P, Weisshaar B, Schmidt T. 2016.** Repeat composition of CenH3-chromatin and H3K9me2-marked heterochromatin in sugar beet (*Beta vulgaris*). *Bmc Plant Biology* **16**: 120.

Kumke K, Macas J, Fuchs J, Altschmied L, Kour J, Dhar MK, Houben A. 2016. *Plantago lagopus* B chromosome is enriched in 5S rDNA-derived satellite DNA. *Cytogenetic and Genome Research* **148**: 68-73.

Lee HR, Zhang WL, Langdon T, Jin WW, Yan HH, Cheng ZK, Jiang JM. 2005. Chromatin immunoprecipitation cloning reveals rapid evolutionary patterns of centromeric DNA in *Oryza* species. *Proc. Nat. Acad. Sci. USA* **102**: 11793-11798.

Li H, Durbin R. 2009. Fast and accurate short read alignment with Burrows-Wheeler transform.
Bioinformatics **25**: 1754-1760.

Liu YL, Su HD, Pang JL, Goo Z, Wang XJ, Birchler JA, Han FP. 2015. Sequential de novo centromere formation and inactivation on a chromosomal fragment in maize. *Proc. Nat. Acad. Sci. USA* **112**: E1263-E1271.

Lu F, Lipka AE, Glaubitz J, Elshire R, Cherney JH, Casler MD, Buckler ES, Costich DE. 2013. Switchgrass genomic diversity, ploidy, and evolution: Novel insights from a network-based SNP discovery protocol. *Plos Genetics* **9**: e1003215.

Macas J, Neumann P, Navrátilová A. 2007. Repetitive DNA in the pea (*Pisum sativum* L.) genome: comprehensive characterization using 454 sequencing and comparison to soybean and *Medicago truncatula*. *BMC Genomics* **8**: 427.

Magoc T, Salzberg SL. 2011. FLASH: fast length adjustment of short reads to improve genome assemblies. *Bioinformatics* **27**: 2957-2963.

Maluszynska J, Heslop-Harrison JS. 1991. Localization of Tandemly Repeated DNA- Sequences in Arabidopsis thaliana.
Plant Journal **1**(2): 159-166.

Melters DP, Bradnam KR, Young HA, Telis N, May MR, Ruby JG, Sebra R, Peluso P, Eid J, Rank D, et al. 2013. Comparative analysis of

tandem repeats from hundreds of species reveals unique insights into centromere evolution. *Genome Biology* **14**: R10.

Miga KH, Newton Y, Jain M, Altemose N, Willard HF, Kent WJ. 2014. Centromere reference models for human chromosomes X and Y satellite arrays. *Genome Research* **24**(4): 697-707.

Murata M, Ogura Y, Motoyoshi F. 1994. Centromeric repetitive sequences in *Arabidopsis thaliana*. *Japanese Journal of Genetics* **69**: 361-370.

Nagaki K, Cheng ZK, Ouyang S, Talbert PB, Kim M, Jones KM, Henikoff S, Buell CR, Jiang JM. 2004. Sequencing of a rice centromere uncovers active genes. *Nat. Genet.* **36**: 138-145.

Nagaki K, Shibata F, Kanatani A, Kashihara K, Murata M. 2012. Isolation of centromeric- tandem repetitive DNA sequences by chromatin affinity purification using a HaloTag7- fused centromere-specific histone H3 in tobacco. *Plant Cell Reports* **31**: 771-779.

Nagaki K, Talbert PB, Zhong CX, Dawe RK, Henikoff S, Jiang JM. 2003. Chromatin immunoprecipitation reveals that the 180-bp satellite repeat is the key functional DNA element of *Arabidopsis thaliana* centromeres. *Genetics* **163**: 1221-1225.

Nasuda S, Hudakova S, Schubert I, Houben A, Endo TR. 2005. Stable barley chromosomes without centromeric repeats. *Proc. Natl. Acad. Sci. USA* **102**: 9842-9847.

Neumann P, Navratilova A, Schroeder-Reiter E, Koblizkova A, Steinbauerova V, Chocholova E, Novak P, Wanner G, Macas J. 2012. Stretching the rules: monocentric chromosomes with multiple centromere domains. *Plos Genetics* **8**: e1002777.

Novak P, Neumann P, Macas J. 2010. Graph-based clustering and characterization of repetitive sequences in next-generation sequencing data. *BMC Bioinformatics* **11**: 378.

Novak P, Neumann P, Pech J, Steinhaisl J, Macas J. 2013. RepeatExplorer: a Galaxy-based web server for genome-wide characterization of eukaryotic repetitive elements from next-generation sequence reads. *Bioinformatics* **29**: 792-793.

Purgato S, Belloni E, Piras FM, Zoli M, Badiale C, Cerutti F, Mazzagatti A, Perini G, Della Valle G, Nergadze SG, et al. 2015. Centromere sliding on a mammalian chromosome. *Chromosoma* **124**: 277-287.

Schueler MG, Higgins AW, Rudd MK, Gustashaw K, Willard HF. 2001. Genomic and genetic definition of a functional human centromere. *Science* **294**(5540): 109-115.

Segal E, Fondufe-Mittendorf Y, Chen LY, Thastrom A, Field Y, Moore IK, Wang JPZ, Widom J. 2006. A genomic code for nucleosome positioning. *Nature* **442**(7104): 772- 778.

Shang WH, Hori T, Toyoda A, Kato J, Popendorf K, Sakakibara Y, Fujiyama A, Fukagawa T. 2010. Chickens possess centromeres with both extended tandem repeats and short non-tandem-repetitive sequences. *Genome Res.* **20**: 1219-1228.

Shibata F, Murata M. 2004. Differential localization of the centromere-specific proteins in the major centromeric satellite of *Arabidopsis thaliana*. *Journal of Cell Science* **117**: 2963- 2970.

Sievers F, Wilm A, Dineen D, Gibson TJ, Karplus K, Li WZ, Lopez R, McWilliam H, Remmert M, Soding J, et al. 2011. Fast, scalable generation of high-quality protein multiple sequence alignments using Clustal Omega. *Molecular Systems Biology* **7**: 539.

Szymanski M, Barciszewska MZ, Erdmann VA, Barciszewski J. 2002. 5S ribosomal RNA database. *Nucleic Acids Research* **30**: 176-178.

Topp CN, Okagaki RJ, Melo JR, Kynast RG, Phillips RL, Dawe RK. 2009. Identification of a maize neocentromere in an oat-maize addition line. *Cytogenet. Genome Res.* **124**: 228- 238.

Triplett JK, Wang YJ, Zhong JS, Kellogg EA. 2012. Five nuclear loci resolve the polyploid history of switchgrass (*Panicum virgatum* L.) and relatives. *Plos One* **7**: e38702.

Valouev A, Ichikawa J, Tonthat T, Stuart J, Ranade S, Peckham H, Zeng K, Malek JA, Costa G, McKernan K, et al. 2008. A high-resolution, nucleosome position map of *C. elegans* reveals a lack of universal sequence-dictated positioning. *Genome Res.* **18**: 1051- 1063.

Wang K, Wu YF, Zhang WL, Dawe RK, Jiang JM. 2014. Maize centromeres expand and adopt a uniform size in the genetic background of oat. *Genome Res.* **24**: 107-116.

Willard HF, Waye JS. 1987. Hierarchical order in chromosome-specific human alpha satellite DNA. *Trends Genet.* **3**: 192-198.

Yan HH, Ito H, Nobuta K, Ouyang S, Jin WW, Tian SL, Lu C, Venu RC, Wang GL, Green PJ, et al. 2006. Genomic and genetic characterization of rice Cen3 reveals extensive transcription and evolutionary implications of a complex centromere. *Plant Cell* **18**: 2123-2133.

Yi CD, Wang MS, Jiang W, Wang DR, Zhou Y, Gong ZY, Liang GH, Gu MH. 2015. Isolation and identification of a functional centromere element in the wild rice species *Oryza granulata* with the GG genome. *Journal of Genetics and Genomics* **42**: 699-702.

Yi CD, Zhang WL, Dai XB, Li X, Gong ZY, Zhou Y, Liang GH, Gu MH. 2013. Identification and diversity of functional centromere satellites in the wild rice species *Oryza brachyantha*. *Chromosome Research* **21**: 725-737.

Zhang HQ, Koblizkova A, Wang K, Gong ZY, Oliveira L, Torres GA, Wu YF, Zhang WL, Novak P, Buell CR, et al. 2014. Boom-bust turnovers of megabase-sized centromeric DNA in Solanum species: rapid evolution of DNA sequences associated with centromeres. *Plant Cell* **26**: 1436-1447.

Zhang T, Talbert PB, Zhang WL, Wu YF, Yang ZJ, Henikoff JG, Henikoff S, Jiang JM. 2013. The CentO satellite confers translational and rotational phasing on cenH3 nucleosomes in rice centromeres. *Proc. Nail. Acad. Sci. USA* **110**: E4875-E4883.

Zhang WL, Friebe B, Gill BS, Jiang JM. 2010. Centromere inactivation and epigenetic modifications of a plant chromosome with three functional centromeres. *Chromosoma* **119**: 553-563.

Zhang WL, Lee HR, Koo DH, Jiang JM. 2008. Epigenetic modification of centromeric chromatin: Hypomethylation of DNA sequences in the CENH3-associated chromatin in *Arabidopsis thaliana* and maize. *Plant Cell* **20**: 25-34.

Zhang WL, Wu YF, Schnable JC, Zeng ZX, Freeling M, Crawford GE, Jiang JM. 2012. High-resolution mapping of open chromatin in the rice genome. *Genome Research* **22**: 151-162.

Zhang Z, Schwartz S, Wagner L, Miller, W. 2000. A greedy algorithm for aligning DNA sequences. *Journal of Computational Biology* **7**: 203-214.

Zhao HN, Zhu XB, Wang K, Gent JI, Zhang WL, Dawe RK, Jiang JM. 2016. Gene expression and chromatin modifications associated with maize centromeres. *G3-Genes Genomes Genetics* **6**: 183-192.

Supporting Information

Figure S1. Mapping of cenH3 ChIP-seq reads on all switchgrass chromosomes.

Figure S2. Dual-color FISH of pairs of centromeric repeats.

Figure S3. Distribution of lengths of DNA fragments derived from ChIP-seq and input libraries.

Figure S4. Nucleosomal wrapping associated with six centromeric satellite repeats in switchgrass.

Figure S5. WW dimer periodicity associated with five switchgrass satellite repeats.

Table S1. Primers used to amplify the centromeric repeats in switchgrass.

Table S2. Number of ChIP-seq fragments and Pv repeat monomers used for consensus construction.

Table S3. PacBio reads containing tandem arrays of Pv repeats.

Table S4. Percent identity matrix of multiple sequence alignment in Figure 7.

Figure legends

Figure 1. Immunofluorescence and ChIP-FISH using anti-cenH3 antibodies. (**A**) Immunofluorescence assay of anti-cenH3 antibody on switchgrass (Kanlow) metaphase chromosomes. Signals were observed exclusively in the centromeres of all switchgrass chromosomes (**B**) FISH on switchgrass (Summer) metaphase chromosomes using a DNA probe derived from ChIPed DNA. FISH signals were concentrated in the centromeric regions and varied in sizes and intensities among different centromeres. Bars = 5 □ m.

Figure 2. Mapping of cenH3 ChIP-seq reads on switchgrass chromosomes. *y* axis represents normalized sequence read count ratio between ChIP-seq and input in 50 kb windows. (**A**) Mapping of cenH3 ChIP-seq reads on switchgrass chromosome 2K. Chromosomal region from ~34.4-36.4 Mb indicated by double red arrows was significantly enriched with ChIP-seq reads. This region likely represents the centromere of this chromosome. (**B**) Mapping of cenH3 ChIP- seq reads on switchgrass chromosome 8K. No sequence enrichment was observed on this chromosome.

Figure 3. Identification of repeat clusters enriched in switchgrass centromeres. (**A**) Computation identification of repeat clusters enriched in switchgrass centromeres. Repeat clusters are represented by dots and their positions reflect the genomic abundance of the corresponding repeats (*x* axis) and their enrichment in ChIP-seq data (*y* axis). The *y* axis represents the ratio of ChIP-seq reads to genomic shot gun sequence reads for each repeat cluster. The *x* axis is the genome proportion of the genomic sequence reads for each repeat cluster. Only the top 200 most abundant repeat clusters are shown. Eight repeat clusters with ChIP-seq enrichment >3-fold are in red color and were selected for cloning and FISH confirmation. (**B**) Amplification of eight centromeric DNA fragments from eight computationally identified repeat clusters from the switchgrass genome. White arrowheads indicate the DNA fragments that were cloned and sequenced. Each of these eight DNA fragments represents a Pv repeat.

Figure 4. FISH mapping of centromeric repeats in switchgrass. (**A**) FISH mapping of repeat Pv36. Signals were detected in the centromeres from a single pair of chromosomes. (**B**) FISH mapping of repeat Pv156. Major signals were detected in the centromeres of one pair of chromosomes. Minor signals (arrows) were detected in another pair of centromeres. (**C**) Dual- color FISH of Pv36 (green) and Pv156 (red). Arrows indicate a pair of minor signals from

pv156. (**D**) FISH mapping of repeat Pv1. Approximately half of the chromosomes showed stronger centromeric signals than the rest of the chromosomes. (**E**) FISH mapping of repeat Pv115. Approximately half of the chromosomes showed stronger centromeric signals than the rest of the chromosomes. (**F**) FISH mapping of repeat Pv29. Two chromosomes showed stronger signals than the rest of the chromosomes. (**G**) FISH mapping of repeat Pv118. Two chromosomes showed relatively stronger signals than the rest of the chromosomes. (**H**) FISH mapping of repeat Pv2. Four chromosomes showed stronger signals than the rest of the chromosomes. (**I**) FISH mapping of repeat Pv45. Four chromosomes (arrows) showed relatively stronger signals than the rest of the chromosomes. Note: weak signals derived from repeats Pv1, Pv2, Pv29, Pv45, Pv115 and Pv118 can be observed on all centromeres in the signal channel.
Bars = 5 μm.

Figure 5. FISH mapping of four centromeric repeats in octoploid switchgrass Trailblazer. (**A**) FISH mapping of repeat Pv36. Signals were detected on four chromosomes. (**B**) FISH mapping of repeat Pv156. Four major signals and four minor signals (arrows) were detected on eight chromosomes. (**C**) FISH mapping of repeat Pv1. Approximately half of the chromosomes showed stronger centromeric signals than the rest of the chromosomes. (**D**) FISH mapping of repeat Pv2. Approximately eight chromosomes showed stronger centromeric signals than the rest of the chromosomes. Bars = 5 μm.

Figure 6. Pv156 is identical to a single unit of the 5S ribosomal RNA gene array. (**A**) A diagram of a 5S rDNA array in switchgrass (upper panel). The 365-bp Pv156 repeat (exemplified) is corresponding to a single unit of the array, including 119 bp coding sequence and 207 bp spacer sequence. The sequence corresponding to the ‘coding region’ is from 49 to 167 bp within Pv156.
(**B**) FISH of Pv156 (left, red signals) and a 5S rDNA probe (middle, green signals). Large arrows (right) indicate a pair of major FISH signals. Small

arrows indicate a pair of minor FISH signals. The red and green signals overlapped completely.

Figure 7. Alignment of consensus sequences of the centromeric satellite repeats from switchgrass and other grass species. The conserved ~80-bp motifs were marked by a black line at the top. The WW dinucleotides that showed ~10 bp periodicity in the conserved 80 bp domain

were marked by asterisks. TR_si, TR_pg, TR_ph are putative centromeric satellite repeats from

Setaria italica, *Pennisetum glaucum*, and *Panicum hallii*, respectively.

Figure 8. Nucleosomal wrapping and sequence features associated with two representative centromeric repeats in Switchgrass. (**A, B**) Positions of nucleosomes along a tetramer of Pv1 (A) and Pv115 (B). Top panels show the distribution of midpoints of small fragments (<130 bp) and large fragments (>130 bp) from the ChIP-seq data. Orange line, small fragments; Green line, large fragments. Bottom panels show wrapping of cenH3 and canonical nucleosomes on tetramers of the Pv1 and Pv115 repeats. Each horizontal line represents a nucleosome-protected region. The *x* axes represent the position on the tetramer. The *y* axes present the length of fragment derived from each nucleosome at a specific position. The thickness of a horizontal line represents the abundance of the sequence. Red line, sequences from ChIP library; Blue line, sequences from input library. Orange and yellow bars represent the four copies of each tetramer. Black rectangles mark the conserved 80-bp motif. The sequences from ChIP and input libraries were plotted separately on the right side of Pv1 and left side of Pv115. (**C, D**) WW dimer periodicity in Pv1 (C) and Pv115 (D). WW dimer periodicity were calculated on the conserved 80-bp motifs and the flanking regions. Normalized ratio of each base pair was calculated by divide the observed WW dimer frequencies by expected frequencies. Phasing score is the median of normalized ratio at WW dimer peaks in Pv1 or Pv115. Red line, WW dimer periodicity associated with the conserved 80 bp motif. Black line, WW dimer periodicity associated with the flanking region. Blue dashed line represents the ~10 bp periodicity.

Table 1. Characteristics of repeat sequence clusters containing centromeric satellite repeats

Cluster	Proportion (%)		Ratio (ChIP-seq/ shotgun)	Repeat	Length of monomer (bp)
	Shotgun read	ChIP-seq read			
CL1	1.62%	6.46%	3.98	Pv 1	166
CL2	0.57%	2.28%	3.99	Pv 2	175
CL29	0.23%	0.69%	3.05	Pv29	175
CL36	0.10%	0.34%	3.50	Pv36	156
CL45	0.17%	0.67%	4.07	Pv45	175
CL115	0.14%	0.53%	3.81	Pv115	166
CL118	0.15%	0.70%	4.59	Pv118	187
CL156	0.24%	5.86%	24.82	Pv156	326

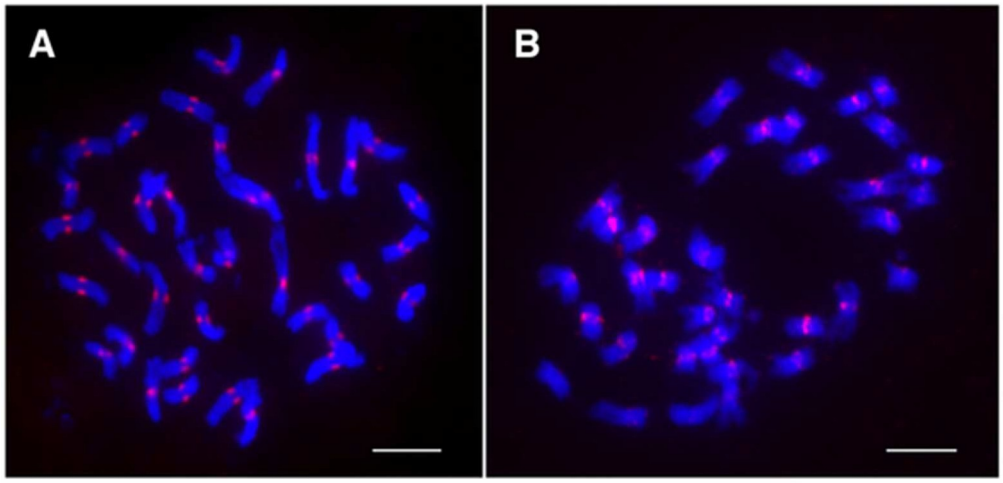

Figure 1

Figure 1. Immunofluorescence and ChIP-FISH using anti-cenH3 antibodies. (A) Immunofluorescence assay of anti-cenH3 antibody on switchgrass (Kanlow) metaphase chromosomes. Signals were observed exclusively in the centromeres of all switchgrass chromosomes (B) FISH on switchgrass (Summer) metaphase chromosomes using a DNA probe derived from ChIPed DNA. FISH signals were concentrated in the centromeric regions and varied in sizes and intensities among different centromeres. Bars = 5 µm.

49x26mm (300 x 300 DPI)

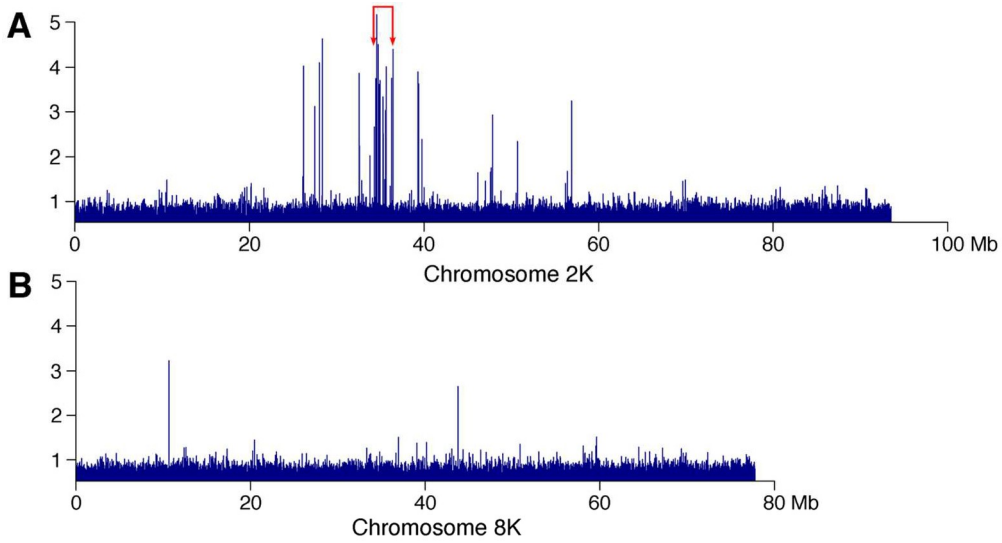

Figure 2

Figure 2. Mapping of cenH3 ChIP-seq reads on switchgrass chromosomes. y axis represents normalized sequence read count ratio between ChIP-seq and input in 50 kb windows. (A) Mapping of cenH3 ChIP-seq reads on switchgrass chromosome 2K. Chromosomal region from ~34.4-36.4 Mb indicated by double red arrows was significantly enriched with ChIP-seq reads. This region likely represents the centromere of this chromosome. (B) Mapping of cenH3 ChIP-seq reads on switchgrass chromosome 8K. No sequence enrichment was observed on this chromosome.

60x36mm (600 x 600 DPI)

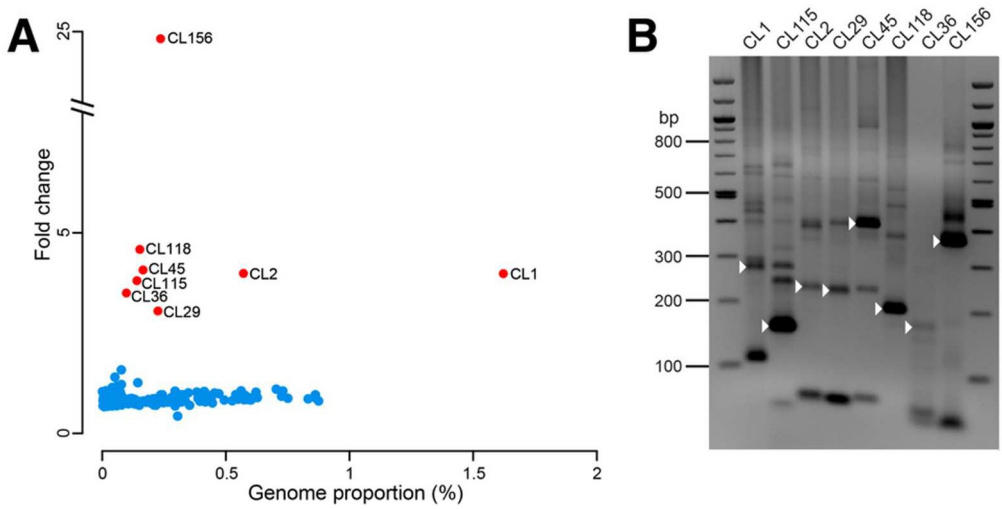

Figure 3

Figure 3. Identification of repeat clusters enriched in switchgrass centromeres. (A) Computation identification of repeat clusters enriched in switchgrass centromeres. Repeat clusters are represented by dots and their positions reflect the genomic abundance of the corresponding repeats (x axis) and their enrichment in ChIP-seq data (y axis). The y axis represents the ratio of ChIP-seq reads to genomic shot gun sequence reads for each repeat cluster. The x axis is the genome proportion of the genomic sequence reads for each repeat cluster. Only the top 200 most abundant repeat clusters are shown. Eight repeat clusters with ChIP-seq enrichment >3-fold are in red color and were selected for cloning and FISH confirmation. (B) Amplification of eight centromeric DNA fragments from eight computationally identified repeat clusters from the switchgrass genome. White arrowheads indicate the DNA fragments that were cloned and sequenced. Each of these eight DNA fragments represents a Pv repeat.

81x48mm (300 x 300 DPI)

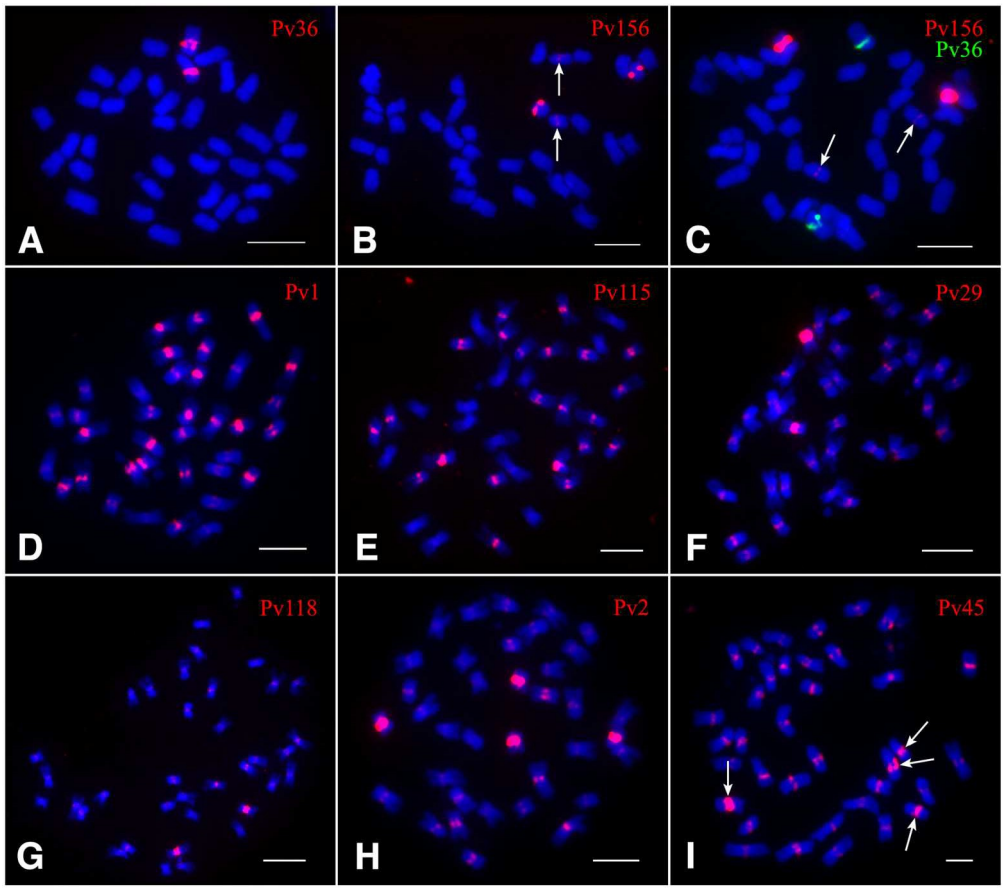

Figure 4

Figure 4. FISH mapping of centromeric repeats in switchgrass. (A) FISH mapping of repeat Pv36. Signals were detected in the centromeres from a single pair of chromosomes. (B) FISH mapping of repeat Pv156. Major signals were detected in the centromeres of one pair of chromosomes. Minor signals (arrows) were detected in another pair of centromeres. (C) Dual-color FISH of Pv36 (green) and Pv156 (red). Arrows indicate a pair of minor signals from pv156. (D) FISH mapping of repeat Pv1. Approximately half of the chromosomes showed stronger centromeric signals than the rest of the chromosomes. (E) FISH mapping of repeat Pv115. Approximately half of the chromosomes showed stronger centromeric signals than the rest of the chromosomes. (F) FISH mapping of repeat Pv29. Two chromosomes showed stronger signals than the rest of the chromosomes. (G) FISH mapping of repeat Pv118. Two chromosomes showed relatively stronger signals than the rest of the chromosomes. (H) FISH mapping of repeat Pv2. Four chromosomes showed stronger signals than the rest of the chromosomes. (I) FISH mapping of repeat Pv45. Four chromosomes (arrows) showed relatively stronger signals than the rest of the chromosomes. Note: weak signals derived from repeats Pv1, Pv2, Pv29, Pv45, Pv115 and Pv118 can be observed on all centromeres in the signal channel. Bars = 5 μm

146x139mm (300 x 300 DPI)

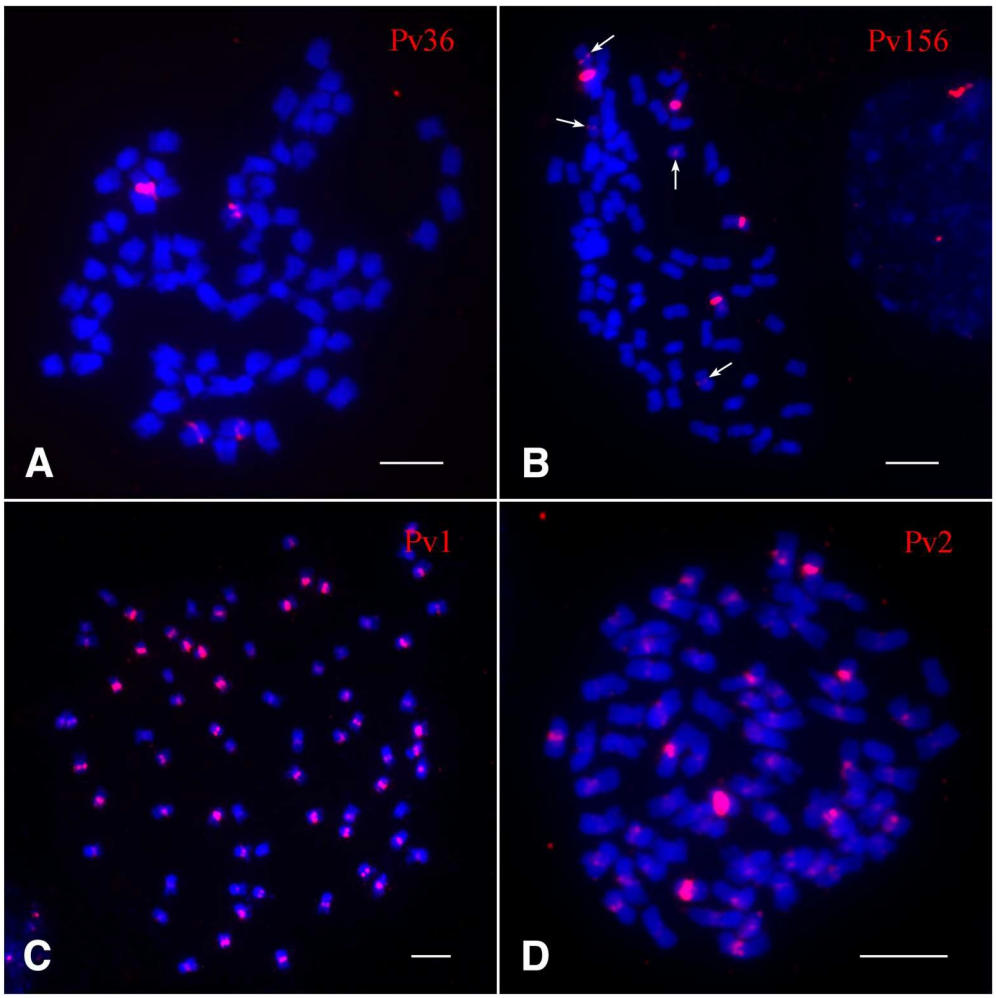

Figure 5

Figure 5. FISH mapping of four centromeric repeats in octoploid switchgrass Trailblazer. (A) FISH mapping of repeat Pv36. Signals were detected on four chromosomes. (B) FISH mapping of repeat Pv156. Four major signals and four minor signals (arrows) were detected on eight chromosomes. (C) FISH mapping of repeat Pv1. Approximately half of the chromosomes showed stronger centromeric signals than the rest of the chromosomes. (D) FISH mapping of repeat Pv2. Approximately eight chromosomes showed stronger centromeric signals than the rest of the chromosomes. Bars = 5 μm

165x176mm (300 x 300 DPI)

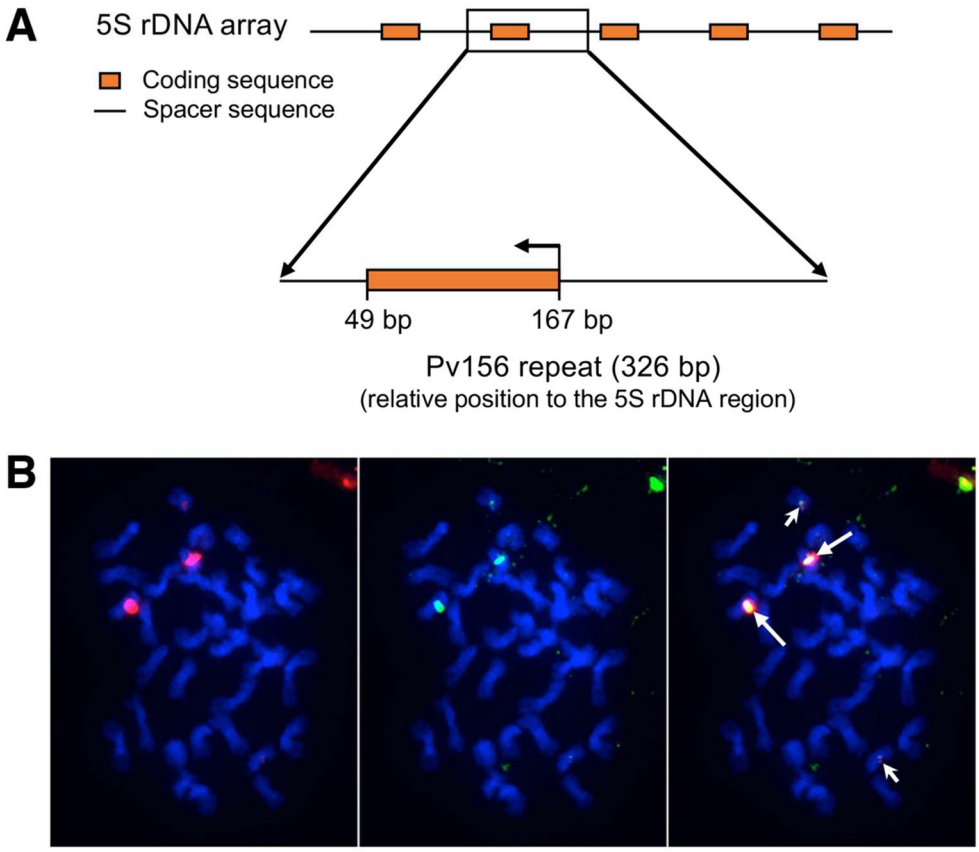

Figure 6

Figure 6. Pv156 is identical to a single unit of the 5S ribosomal RNA gene array. (A) A diagram of a 5S rDNA array in switchgrass (upper panel). The 365-bp Pv156 repeat (exemplified) is corresponding to a single unit of the array, including 119 bp coding sequence and 207 bp spacer sequence. The sequence corresponding to the 'coding region' is from 49 to 167 bp within Pv156. (B) FISH of Pv156 (left, red signals) and a 5S rDNA probe (middle, green signals). Large arrows (right) indicate a pair of major FISH signals. Small arrows indicate a pair of minor FISH signals. The red and green signals overlapped completely.

98x94mm (300 x 300 DPI)

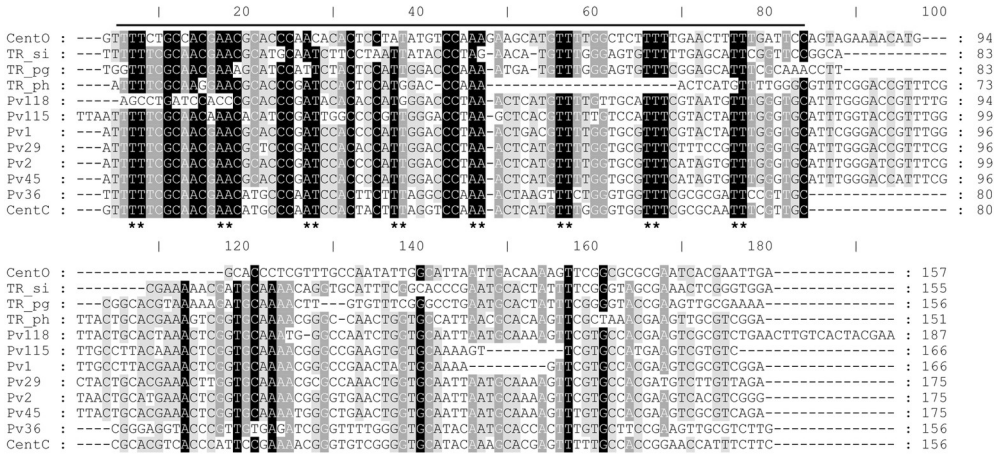

Figure 7

Figure 7. Alignment of consensus sequences of the centromeric satellite repeats from switchgrass and other grass species. The conserved ~80-bp motifs were marked by a black line at the top. The WW dinucleotides that showed ~10 bp periodicity in the conserved 80 bp domain were marked by asterisks. TR_si, TR_pg, TR_ph are putative centromeric satellite repeats from Setaria italica, Pennisetum glaucum, and Panicum hallii, respectively.

73x37mm (600 x 600 DPI)

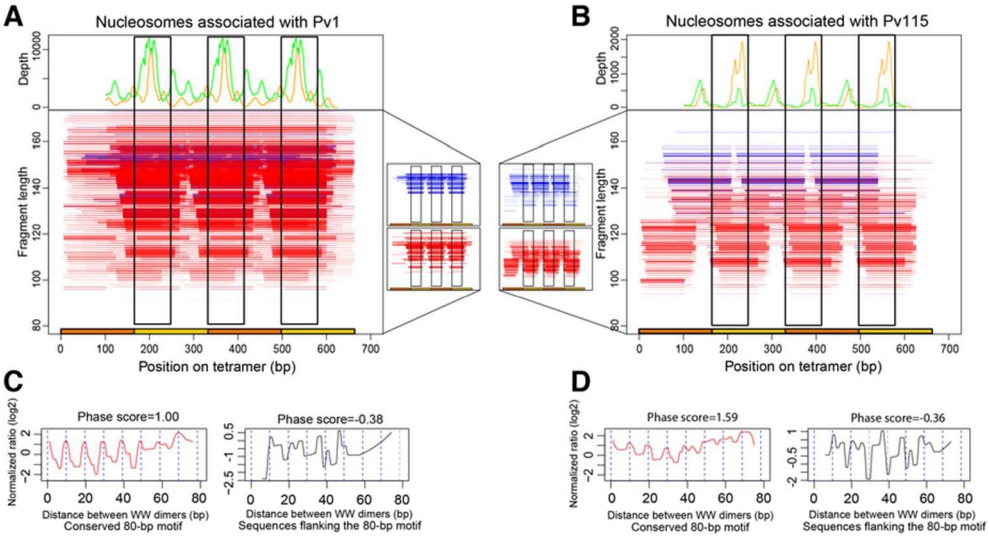

Figure 8

Figure 8. Nucleosomal wrapping and sequence features associated with two representative centromeric repeats in Switchgrass. (A, B) Positions of nucleosomes along a tetramer of Pv1 (A) and Pv115 (B). Top panels show the distribution of midpoints of small fragments (<130 bp) and large fragments (>130 bp) from the ChIP-seq data. Orange line, small fragments; Green line, large fragments. Bottom panels show wrapping of cenH3 and canonical nucleosomes on tetramers of the Pv1 and Pv115 repeats. Each horizontal line represents a nucleosome-protected region. The x axes represent the position on the tetramer. The y axes present the length of fragment derived from each nucleosome at a specific position. The thickness of a horizontal line represents the abundance of the sequence. Red line, sequences from ChIP library; Blue line, sequences from input library. Orange and yellow bars represent the four copies of each tetramer. Black rectangles mark the conserved 80-bp motif. The sequences from ChIP and input libraries were plotted separately on the right side of Pv1 and left side of Pv115. (C, D) WW dimer periodicity in Pv1 (C) and Pv115 (D). WW dimer periodicity were calculated on the conserved 80-bp motifs and the flanking regions. Normalized ratio of each base pair was calculated by divide the observed WW dimer frequencies by expected frequencies. Phasing score is the median of normalized ratio at WW dimer peaks in Pv1 or Pv115. Red line, WW dimer periodicity associated with the conserved 80 bp motif. Black line, WW dimer periodicity associated with the flanking region. Blue dashed line represents the ~10 bp periodicity.

79x47mm (300 x 300 DPI)